DOCUMENT RESUME

ED 327 546 TM 015 032

AUTHOR Sykes, Robert C.
TITLE Stability of IRT b-Values over Time and Position.
PUB DATE 89
NOTE 30p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Covariance; Certification; Comparative Testing; *Item Response Theory; *Licensing Examinations (Professions); Predictive Measurement; *Research Methodology; *Test Items
IDENTIFIERS *B Values; *Invariance Principle; Item Position (Tests); Rasch Model

ABSTRACT

An analysis-of-covariance methodology was used to investigate whether there were population differences between tryout and operational Rasch item b-values relative to differences between pairs of item response theory (IRT) b-values from consecutive operational item administrations. This methodology allowed the evaluation of whether any such differences could be attributed to differences in item position after controlling for the background effect of scale drift on b-values. Usable real or scored items in a large licensure examination were selected for study. The final sample consisted of 406 administration difference pairs. For each item, a record was created containing: (1) administration dates; (2) item positions for several administrations ending with July 1987; (3) classification code; (4) type of administration (real or tryout); and (5) b-value and standard error. The block-wise placement of tryout items had no apparent effect on their b-values. Neither differences in booklet nor test item locations were significantly associated with differences in b-values across consecutive pairs of administrations. A small but significant increase in b-value differences in pairs having earlier administration (between February 1981 and July 1986) suggested that scale drift increased over this period. Three data tables and eight figures are included. (Author/SLD)

Stability of IRT b-values over Time and Position

Robert C. Sykes

CTB/MacMillan/McGraw-Hill

Introduction

The use of item preequating to calibrate tryout items onto an IRT scale provides a means to build large pools of calibrated items. The grouping of these tryout items as blocks of consecutive items affords cost efficiencies in the printing and scoring of test booklets, but raises questions concerning whether the restricted placement of tryout items relative to operational items induces context or order effects. An Analysis of Covariance methodology was employed in this study to investigate whether there were population differences between tryout and operational Rasch item b-values relative to differences between pairs of b-values from consecutive operational item administrations. The Analysis of Covariance methodology allowed an evaluation of whether any such differences could be attributed to differences in item position after controlling for the background effect of scale drift on b-values.

As noted by Wainer and Kiely (1987) in their discussion of the potential impact of context effects on CAT (Computerized Adaptive Testing) scores or item parameters, the influence that item location within a test may have on item parameter estimates is one type of undesirable item context effect. Several studies have documented differences in item difficulty parameter estimates as a function of their location Whitely and Dawis (1976) reported that for 9 out of 15 verbal analogy items embedded in seven different unspeeded tests parameter estimates differed when parameters were obtained for each test. Yen (1980) found that for reading comprehension items administered in a nonspeeded test, items became more difficult the later in the test they appeared. A preequating study by Eigner and Cook (1983) replicated Yen's findings. Wise et al (1989) also found item position effects for Word Knowledge and Arithmetic Reasoning tests.

If the effects of differences in item location are assessed on parameter estimates obtained from item administrations spanning a length of time, estimates of such differences may be biased by the presence of background scale drift. Bock, Muraki, and Pfeiffenberger (1988) document the presence of differential linear drift of location parameters of College Board Physics Achievement items over a 10-year period. Additionally, they present statistical procedures which may be used for detecting and estimating item parameter drift in item pools.

The potential presence of scale drift in long-term testing programs would thus necessitate that any evaluation of change in item parameter estimates incorporate procedures for controlling, and perhaps estimating the magnitude of the drift if parameter estimates are not obtained from concurrent item administrations. Incorporation, in a design methodology, of statistical procedures for controlling the effect of drift allows independent estimates of the mean difference in item parameters across type of administration (tryout vs operational) as well as an independent

evaluation of the extent to which such changes may be attributed to differences in item location across test administrations.

Method

All usable real or scored items in a large licensure examination were selected for study. A predominant number of the test items for the examination are typically associated with passages. For each of the real or scored items in the four booklet test, item statistics for a 7/87 administration as well as all previous administrations were obtained. For each administration of each item, a record was created containing:

1) Administration date
2) Classification code
3) Position in the test booklet and complete exam indices
4) Type of administration, real or tryout, and
5) b-value and standard error.

The selected 7/87 form had been equated to previous forms by setting the mean 7/87 b-value, obtained from all scored items, equal to the mean of the b-values obtained from the last administration of the scored items.

The presence of item statistics or parameters from multiple administrations of each item implied that any assessed item statistic would not be independent across the observational units, that is, the test administrations. In order to obtain an independent unit of analysis, differences in b-values for all pairs of consecutive administrations (later administration minus earlier administration) were computed. Thus, if an item was administered on three other occasions besides 7/87, say 2/86, 7/84, and 7/82, three b-value differences would be generated: the b-value in 7/87 minus the b-value obtained in 2/86, the 2/86 b-value minus the 7/84 b-value, and the 7/84 b-value minus the 7/82 b-value.

Independent differences in b-values across pairs of consecutive administrations permitted an evaluation of the stability of b-values using an Analysis of Covariance methodology. This methodology provided a means to:

1) Determine the presence of significant linear relationships between differences in item position in the test booklet or complete exam and changes in b-values across administrations and

2) Assess possible "background" scale drift confounding differences in b-values.

Furthermore, the Analysis of Covariance promised more powerful significance tests of differences between tryout and real item statistics by regressing error on any position or time index having a significant relationship with b-value differences.

The sample of paired, consecutive administration b-value differences was used for a set of comparative analyses. In the first analysis, differences in administration b-values (noted in tables and figures as BVDif) were assessed in a crossed, unbalanced design consisting of three classification variables:

1) Type of administration pair (TypePr): both real administrations vs. a later real, earlier tryout administration.

2) Domain 1 (NP) classification of the item (5 classes) and all its consecutive administration pairs, and

3) Domain 2 (CN) classification of the item (4 classes) and all its consecutive administration pairs.

Paired administration b-value differences were compared across levels of the classification variables after fitting the most parsimonious model for the linear regression of the errors on the covariables.

The initial regression model consisted of four covariables. Difference in book position (BP1-BP2) and difference in position in the complete exam (TP1-TP2) were entered last and assessed for significance. Given the length of the exam, four test booklets of 93 items given over the span of two days, it was conceivable that fatigue or motivational effects may be more substantial near the end of the test than near the end of the first booklets. The magnitude of associations between the two position indices and b-value differences might then be expected to reflect these differences in effects. For both indices position in the second, earlier administration was subtracted from the position in the first, later administration.

The two other covariables served as controls for potential scale drift and consequently were entered before the difference-in-position indices. The first of these control covariables was an index of time elapsed, in months, from 2/81 (TimFr281) to the earliest administration in each consecutive administration pair. This covariable would demonstrate a significant positive correlation with the b-value differences if scale drift had occurred over the span of six and a half years, 2/81 through

7/87, and if the drift consisted of an increase or decrease in mean item difficulty or candidate capability at several occasions over this period. Negative equating constants for every administration back to at least 2/85 suggested a shift in the mean candidate capability during this period.

The second of the control covariables was time elapsed (in months) between the earlier and the later administrations in each administration pair (TimBtwAd). The purpose of this index was to ensure that any difference between real-real administration b-value differences and real-tryout administration b-value differences could not be attributed to differences in the elapsed time between the two types of administration pairs. The shorter, on average, time elapsing before a tryout can be administered as a real compared to the average time a real item is retired before readministration might be expected to produce smaller mean differences for real-tryout administration pairs if significant scale drift occurs across the period spanned.

Analyses and Results

Analysis of b-value differences

Initial plots of b-value differences by the four initial covariables were scanned for outliers. Six consecutive administration pairs were deleted because of extreme b-value differences. A total of 406 administration difference pairs remained in the sample.

Means and standard deviations of the dependent variable and four prospective covariates are presented in Table 1. In addition to the BVDif index of change in b-values, the difference in consecutive administration b-values relative to the standard error of the later administration (BVDif/s.e.1) is provided, hereafter referred to as BVDifSt.

The average b-value difference of .01 does not differ significantly from 0 when assessed against its standard error (.01). B-value differences are, on average, .29 of a real item standard error. This is slightly more than 1.6 times the standard error of the BVDifSt mean (.18) as opposed to the one standard error of the BVDif mean represented by the average BVDif.

The average time elapsed between the later (chronologically) real administration and the earlier real or tryout administration in the paired consecutive administrations was 22.99 months. The average difference in test booklet position and test position between the later and earlier administrations was negative for both position indices, -23.78 and -58.37, respectively. Both mean BP1-BP2 and mean TP1-TP2 are substantially less than 0

because of the substantial number of tryout administrations. Their higher (i.e., later on average) booklet placement and higher complete test position are subtracted from lower, on average, book and test positions for the next real administration.

Turning to the correlations and their significance levels (sign.) also presented in Table 1, a small but significant positive correlation of .13 is found between the b-value differences and elapsed time from 2/81. B-values of the later administration tend to increase relative to the earlier administration as time from 2/81 increases. The positive association between the b-value differences and elapsed time from 2/81 is slightly more pronounced when the differences are represented relative to a standard error, viz. a significant .16 correlation between BVDifSt and TimFr281. A significant positive association between these two indices might be expected in the presence of a scale drift that increased between 2/81 and 7/87.

The significant associations demonstrated between TimFr281 and both TimBtwAd and TP1-TP2, as well as the significant, smaller in absolute value, negative association between TP1-TP2 and TimBtwAd are artifacts. For example, a substantial negative correlation of -.51 was observed between TimFr281 and TimBtwAd because the longer the time elapsed from 2/81, the shorter the maximum possible time between administrations, given the upper bound of the 7/87 administration date.

The differences in consecutive administration b-values, less the six paired observations deleted as outliers, were plotted against the four prospecti covariates for signs of nonlinearity between the dependent variable and the covariates. These plots may be found in Figures 1-6b. Figures 1 and 2 provide no indication that b-value differences across groups or classification cells are nonlinearly related to differences in test position (TP1-TP2) or differences in book position (BP1-BP2), respectively. An examination of plots of b-value differences and differences in book position within type of administration pair in Figures 3a and 3b also revealed no sign of nonlinearity and similar mean differences of approximately 0.

The plots of b-value differences against the elapsed time indices, TimBtwAd and TimFr281, in Figures 4 through 6b also indicate no sign of nonlinearity. The marginal distributions of b-value differences within type of administration pair in Figures 6a and 6b appear homogenous.

<u>Model-fitting</u>

The significance of the linear associations between the four prospective covariates--TimFr281, TimBtAd, BP1-BP2, and TP1-TP2--and b-value differences were evaluated in a step-wise Analysis of

Regression. In order to perform a statistical test of parallelism of regression slopes, it was necessary to initially simplify the model. A multiple regression was performed controlling for any group effects. These results are presented in Table 2.

Two covariates, TimBtAd and TP1-TP2, were eliminated because of nonsignificant linear relationships with b-value differences. When entered last in the regression, Fs' of .12 and .00 were obtained, with respective p values of .73 and .96. After elimination from the model, the slopes of the b-value differences regressed on the two remaining covariates did not differ significantly across groups.

Of the two remaining covariates, TimFr281 and BP1-BP2, neither explained significant variation in b-value differences when entered last though TimFr281 was marginally insignificant ($p = .06$ and $p = .96$, respectively, in Table 2). TimFr281 was significant when entered first ($F = 5.94$, $p = .02$).

Given the significance of the single predictor, TimFr281, the classification effects were next assessed in the Analysis of Covariance summarized in Table 3. As in the Analysis of Regression, terms were evaluated in a step-wise fashion but, for the covariance analysis, the classification variables were assessed after controlling for TimFr281.

Neither the three-way nor any of the three two-way interactions were significant after controlling for lower-order interactions or main effects (no $p < .21$). The effect of type of administration pair, tested only after any systematic differences in b-values due to content classification had been accounted for, was insignificant at $p = .22$. Differences in the mean b-value differences across Domain 2 (CN) levels not attributable to the Domain 1 (NP) classification were significant at $p = .04$. Because NP was the first classification effect entered into the model, the insignificant F (.49) for the Type I sum of squares indicated no significant difference in b-value differences after controlling for TimFR281 but ignoring the other classification effects.

A final model incorporating a single covariate, TimFr281, and the CN classification was fitted. A test of the parallelism of the regression of b-value differences on TimFr281 within each CN classification was insignificant at $P = .48$, indicating the sufficiency of a common regression slope fitted across the four CN levels. The single-factor model with one covariate explained significant variation in BVDif ($p < .005$).

Only one of three estimable CN level effects was significant at $P < .05$: Physiological Integrity. The final fitted model for b-value differences across consecutive administration pairs <u>of Physiological Integrity items</u> was consequently:

$$\text{Est. B-value difference} = -.131 + .082 + .002\ (\text{TimFr281} - 46.12) + e'$$

For the other three CN levels, 0 replaces .082.

<u>Discussion</u>

The assessment of change in item b-values across consecutive pairs of item administrations indicated no signs that the placement of tryout items had a significant effect on the tryout b-values. The absence of linear relationships between the b-value differences and differences in book position or complete test position, coupled with the fact that the test booklets (and complete exam) are unspeeded, suggests that general motivational or fatigue effects, which might influence performance on lengthy exams, are not conspicuously present for tryout <u>or real</u> items. There appeared to be no low-order effect of item position on the predominantly passage-related items. Furthermore, plots of b-value differences provided no indication that there were nonlinear or more localized effects of item placement on either real or tryout b-values.

The positive relationship between time elapsed from 2/81 and b-value differences (i.e., the significant .002 regression coefficient) implies the presence of an <u>increasing</u> drift in the scale in the direction of increasing item difficulty between 2/81 and approximately 7/86, the latest (chronologically) "early" member of a consecutive administration pair. This finding is consistent with the equating constants for RN 2/85 through 7/86 being increasingly negative (-.032. -.040. -.048, and -.060, respectively). A prediction of whether the drift has continued to increase after 7/86 would require extrapolation from the model which is hazardous. Equating constants for the exams after the 7/86 and excluding the 2/88 (-.079, -.46, -.141, and -.069 for 7/86 through 2/89) do not support a trend of invariably more negative equating constants though, even excluding 7/88, the average post 7/86 equating constant (-.065) is less than the average of the four previous exam equating constants (-.045).

Finally, the fitted model for b-value differences imputes a differential scale drift over Domain 2 categories. The average b-value of one category of CN items increased faster than the average item b-value across all content classifications.

Conclusions

1) The blockwise placement of tryout items had no apparent effect on the b-values of the tryout items.

2) Neither differences in booklet or complete exam item locations were significantly associated with differences in b-values across consecutive pairs of administrations. This result is somewhat surprizing given the predominance of test items which were passage related.

3) A small but significant increase in b-value differences in pairs of consecutive administrations having earlier administrations between 2/81 and 7/86 suggests that scale drift had increased over this period.

4) B-values increased to a significantly greater extent for one category of Domain 2 (CN) items than they did, on average, across all content categories.

References

Bock, R.D. (1975). Multivariate statistical methods in behavioral research. New York, NY: McGraw-Hill, 1975.

Bock, R.D, Muraki, E. & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. Journal Educational Measurement, 25, 275-285.

Eignor, D.R., & Cook, L.L. (1983). An investigation of the feasibility of using item response theory in the preequating of aptitude tests. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case of testlets. Journal of Educational Measurement, 24, 185-201.

Whitely, S.E., & Dawis, R.V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 329-337.

Wise, L.L., Chia, W.J. & Park, R.F. (1989). Effects of items position on IRT parameter estimates and item statistics. Paper presented at the annual meeting of the American Educational Research Association.

Yen, W.M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. Journal of Education Measurement, 17, 297-311.

Yen, W.M. (1981). Using simulation results to choose a latent trait model, Applied Psychological Measurement, 2, 245-262.

Table 1

Means, Standard Deviations, and Correlations of the Consecutive Administration Item Statistic Differences (N = 406)

	BVDif	BVDifSt	TimFr281	TimBtwAd	BP1-BP2	TP1-TP2
Mean	.01	.29	46.12	22.99	-23.78	-58.37
sd	.26	3.64	15.11	9.48	32.55	174.93
BVDif	1.00					
sign.						
BVDifSt	.98	1.00				
sign.	.00					
TimFr281	.13	.16	1.00			
sign.	.01	.00				
TimBtwAd	-.07	-.09	-.51	1.00		
sign.	.15	.07	.00			
BP1-BP2	-.02	-.03	.02	-.01	1.00	
sign.	.65	.57	.74	.77		
TP1-TP2	.07	.07	.36	-.19	.03	1.00
sign.	.18	.15	.00	.00	.60	

Table 2

Analysis of Regression
Dependent Variable: b Value Difference

Source	df	Type I ss	F Value	Pr > F	Type III ss	F Value	Pr > F
Regression effects eliminating design effects	4	.434	1.67	> .10			
TimFr281	1	.385	5.94	.02	.225	3.48	.06
TimBtwAd	1	.006	0.09	.76	.008	0.12	.73
BP1-BP2	1	.043	0.66	.42	.043	0.66	.42
TP1-TP2	1	.000	0.00	.96	.000	0.00	.96
Reduced residual	362	23.474					
Error	366	23.908					

Table 3

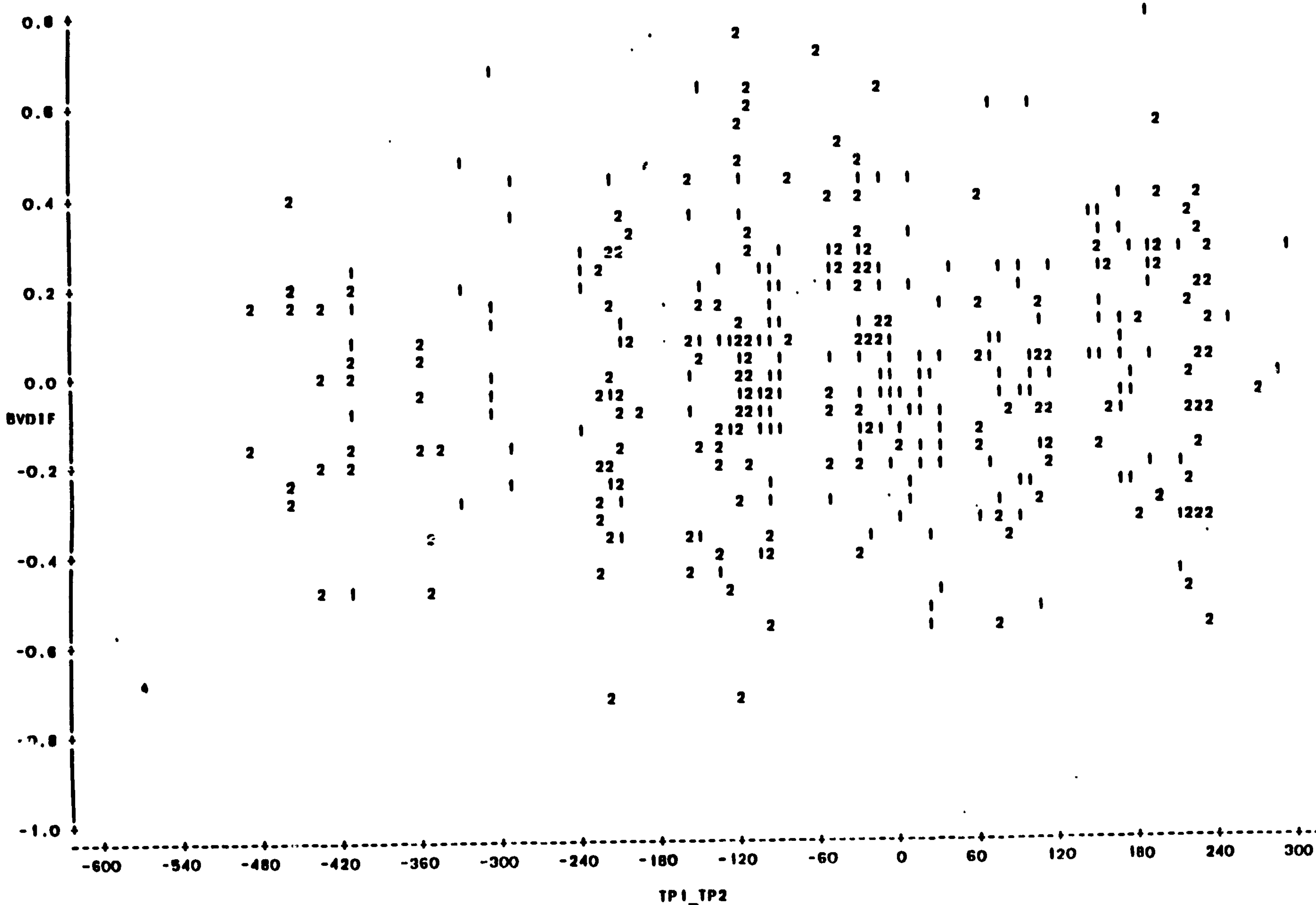
Analysis of Covariance
Dependent Variable: b Value Difference

Source	df	Type I ss	F value	Pr > F
Regression				
Timfr281	1	.444	6.89	.01
NP	4	.126	.49	.74
CN	3	.557	2.88	.04
TypePr	1	.099	1.53	.22
NP * CN	12	.781	1.01	.44
TypePr * NP	4	.277	1.07	.37
TypePr * CN	3	.066	.34	.79
TypePr * NP * CN	12	1.013	1.31	.21
Reduced Residual	365	23.523		
Total (Corrected)	405	26.887		

FIGURE 1

PLOT OF B-VALUE DIFFERENCE BY DIFFERENCE IN TEST POSITION (TP1-TP2)

PLOT OF BVDIF*TP1_TP2 SYMBOL IS VALUE OF TYPEPR

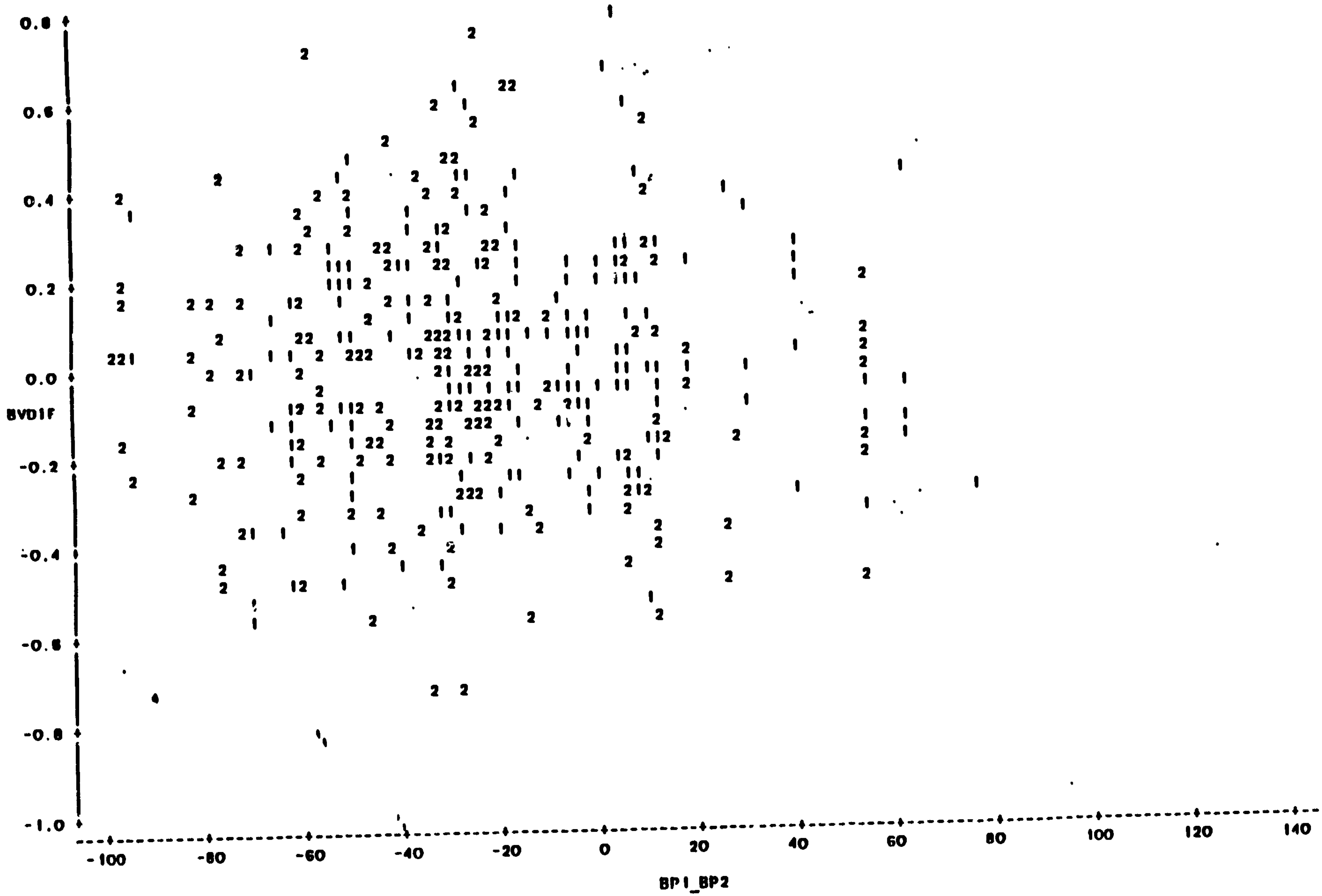


NOTE: 69 OBS HIDDEN

FIGURE 2

PLOT OF B-VALUE DIFFERENCE BY
DIFFERENCE IN TEST BOOKLET POSITION
(BP1-BP2)

PLOT OF BVDIF*BP1_BP2   SYMBOL IS VALUE OF TYPEPR



NOTE: 73 OBS HIDDEN

FIGURE 3a
PLOT OF B-VALUE DIFFERENCE BY
DIFFERENCE IN TEST BOOKLET POSITION
REAL-REAL ADMINISTRATIONS (TypeFr=1)

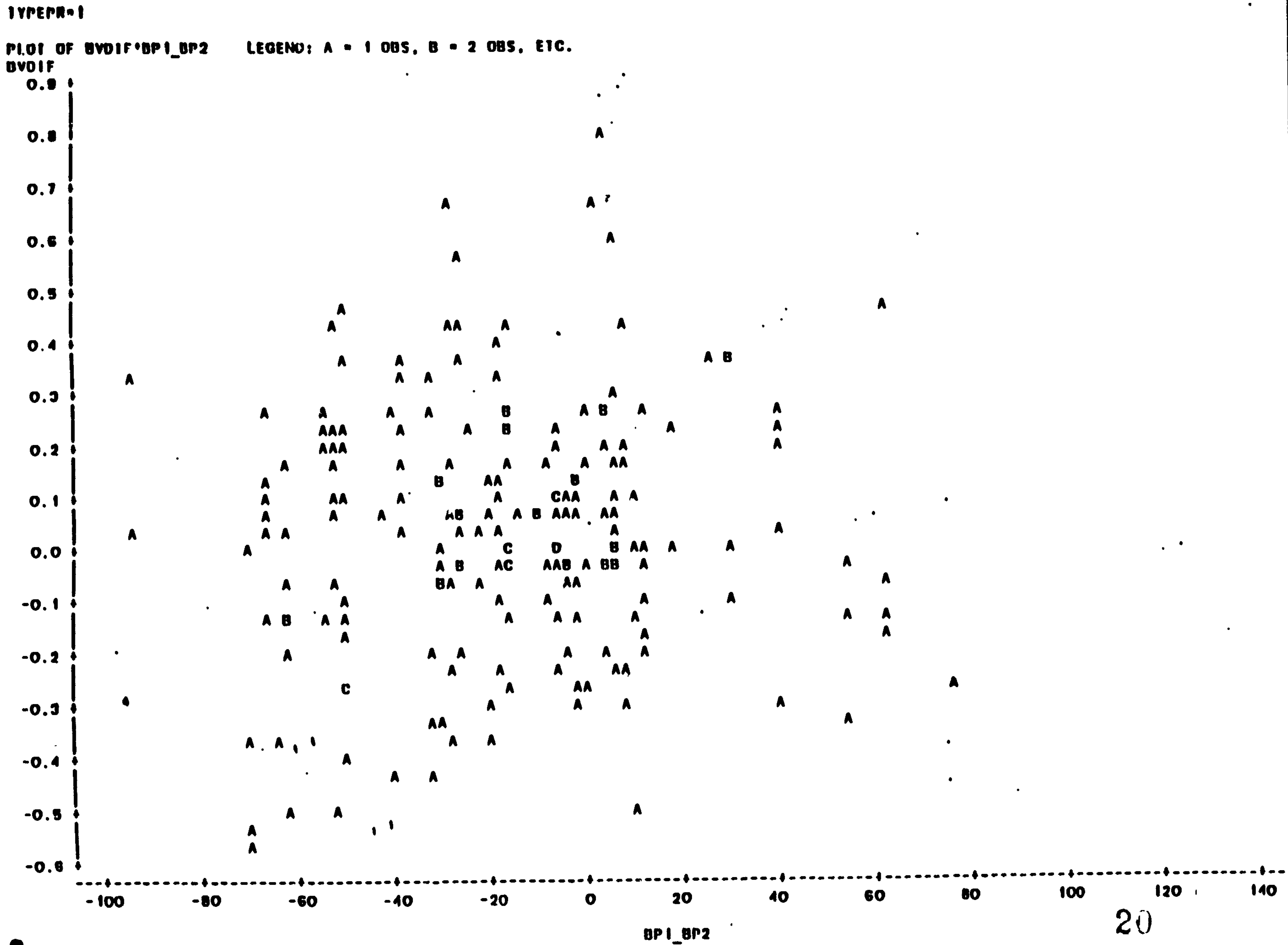


FIGURE 3b
PLOT OF B-VALUE DIFFERENCE BY
DIFFERENCE IN TEST BOOKLET POSITION
REAL-TRYOUT ADMINISTRATIONS (TypePr=2)

TYPEPR=2

PLOT OF BVDIF*BP1_BP2   LEGEND: A = 1 OBS. B = 2 OBS, ETC.

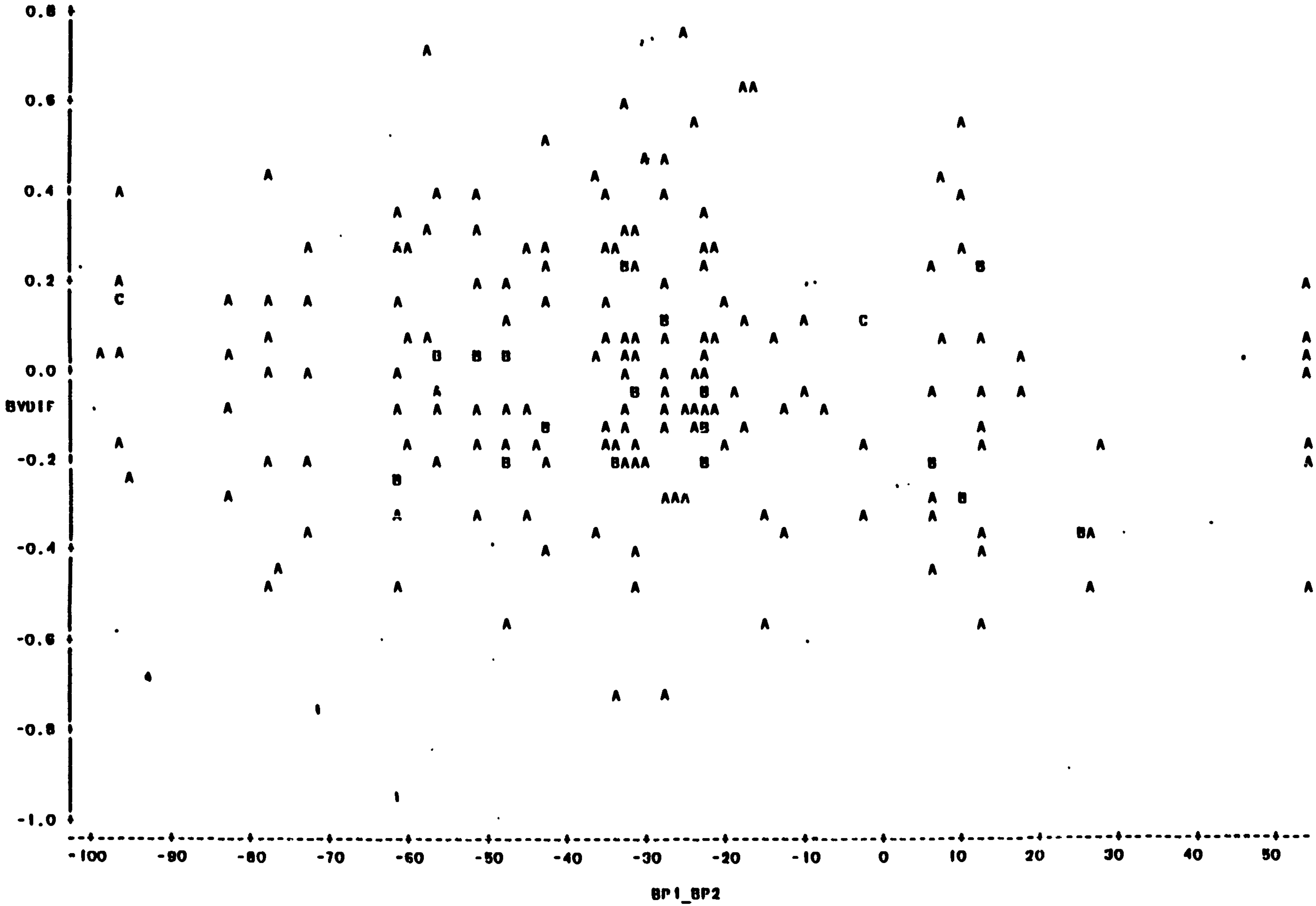
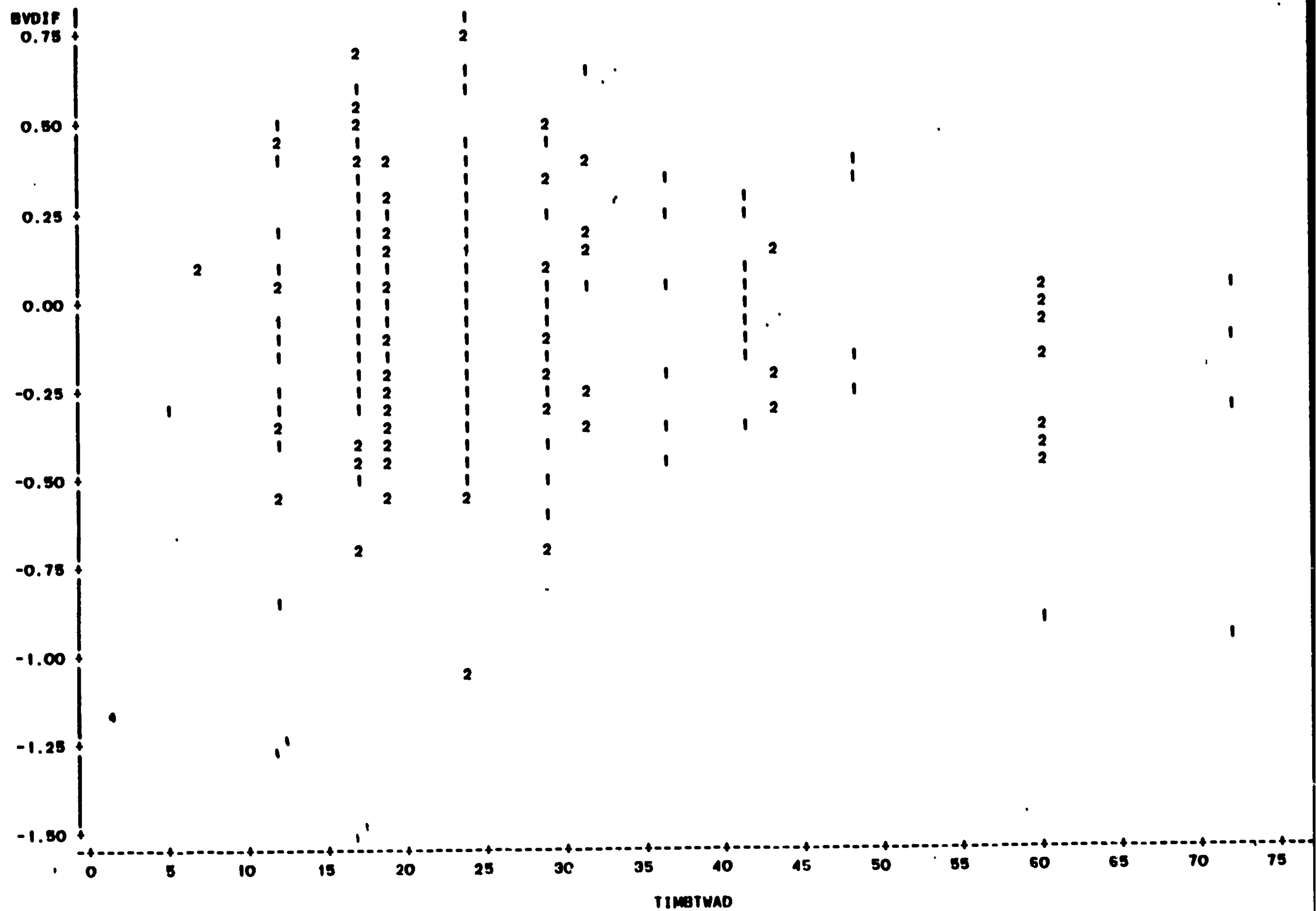


FIGURE 4

PLOT OF B-VALUE DIFFERENCE BY
TIME BETWEEN ADMINISTRATIONS (Timbtwad)

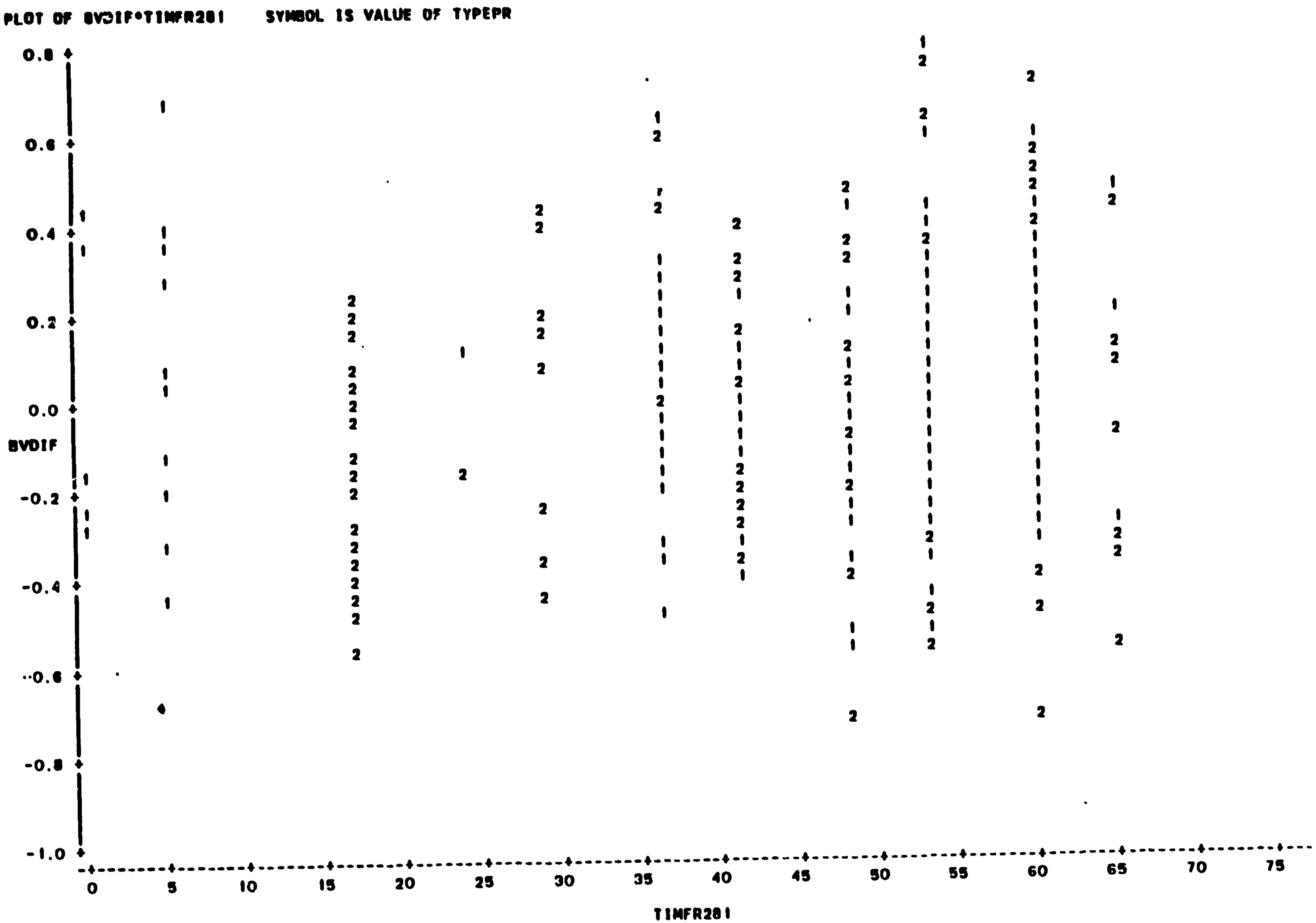
PLOT OF BVDIF*TIMBTWAD   SYMBOL IS VALUE OF TYPEPR



NOTE: 269 OBS HIDDEN

FIGURE 5

PLOT OF B-VALUE DIFFERENCE BY
TIME FROM 2/81 TO EARLIEST ADMINISTRATION (TimFr281)



NOTE: 236 OBS HIDDEN

FIGURE 6a

PLOT OF B-VALUE DIFFERENCE BY

TIME FROM 2/81 TO EARLIEST ADMINISTRATION (TimFr281)

REAL-REAL ADMINISTRATIONS (TypePr=1)

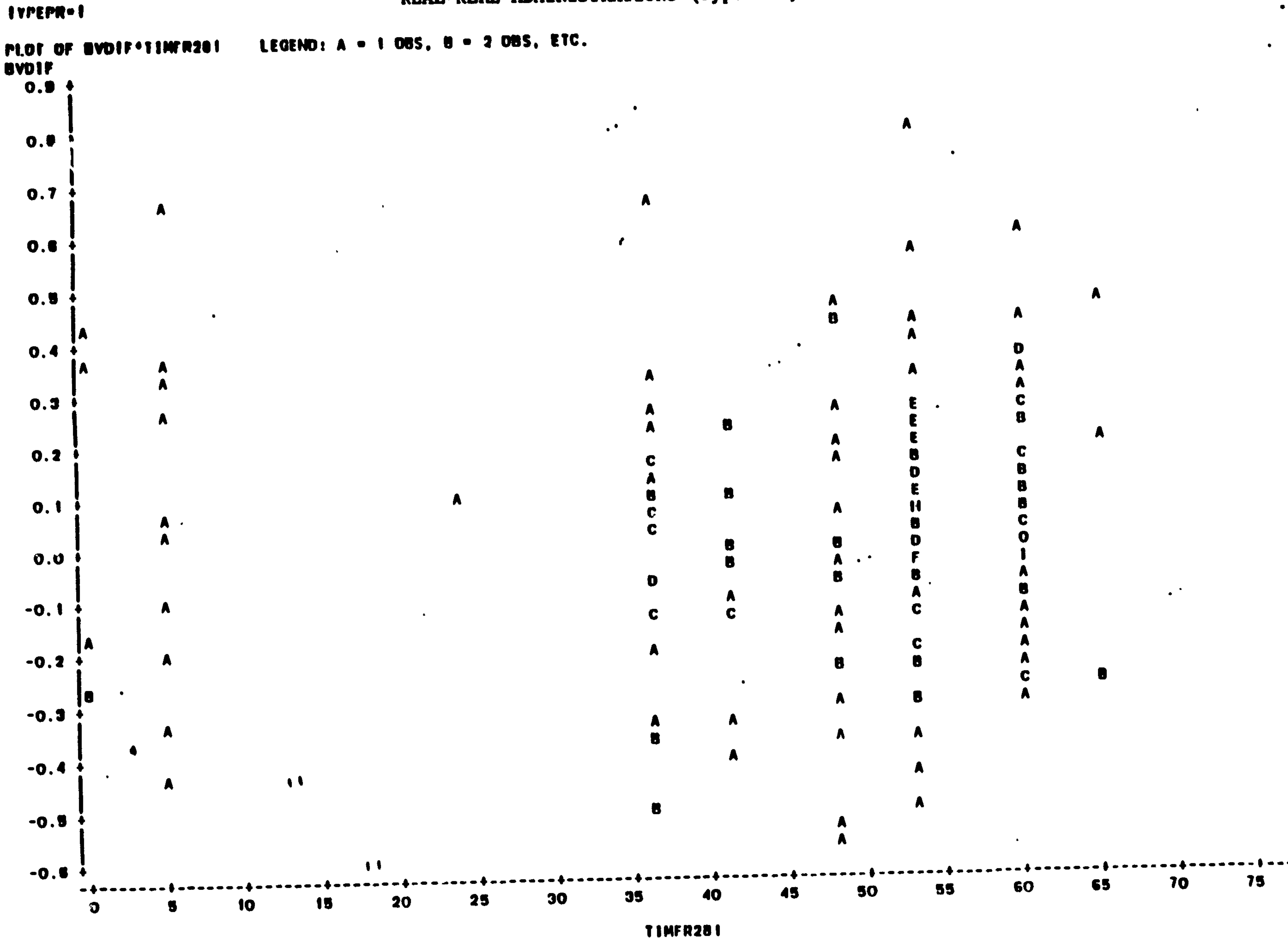


FIGURE 6b

PLOT OF B-VALUE DIFFERENCE BY
TIME FROM 2/81 TO EARLIEST ADMINISTRATION (TimFr281)
REAL-TRYOUT ADMINISTRATIONS (TypePr=2)

TYPEPR=2

PLOT OF BVDIF*TIMFR281 LEGEND: A = 1 OBS, B = 2 OBS, ETC.